



## Article

# Traceback and Chunk-Based Learning: Comparing Usage-Based Computational Approaches to Child Code-Mixing

Nikolas Koch <sup>1</sup>, Stefan Hartmann <sup>2,\*</sup> and Antje Endesfelder Quick <sup>3</sup>

<sup>1</sup> Institute for German as a Foreign Language, Ludwig-Maximilians-Universität Munich, 80539 Munich, Germany

<sup>2</sup> German Department, University of Düsseldorf, 40225 Düsseldorf, Germany

<sup>3</sup> Institute of British Studies, University of Leipzig, 04107 Leipzig, Germany

\* Correspondence: hartmast@hhu.de

**Abstract:** Recent years have seen increased interest in code-mixing from a usage-based perspective. In usage-based approaches to monolingual language acquisition, a number of methods have been developed that allow for detecting patterns from usage data. In this paper, we evaluate two of those methods with regard to their performance when applied to code-mixing data: the traceback method, as well as the chunk-based learner model. Both methods make it possible to automatically detect patterns in speech data. In doing so, however, they place different theoretical emphases: while traceback focuses on frame-and-slot patterns, chunk-based learner focuses on chunking processes. Both methods are applied to the code-mixing of a German–English bilingual child between the ages of 2;3 and 3;11. Advantages and disadvantages of both methods will be discussed, and the results will be interpreted against the background of usage-based approaches.

**Keywords:** code-mixing; usage-based approach; bilingual first language acquisition; traceback; chunk-based learning



**Citation:** Koch, Nikolas, Stefan Hartmann, and Antje Endesfelder Quick. 2022. Traceback and Chunk-Based Learning: Comparing Usage-Based Computational Approaches to Child Code-Mixing. *Languages* 7: 271. <https://doi.org/10.3390/languages7040271>

Academic Editors: A. Seza Doğruöz and Sunayana Sitaram

Received: 18 May 2022

Accepted: 5 September 2022

Published: 25 October 2022

**Publisher's Note:** MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



**Copyright:** © 2022 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

## 1. Introduction

The most salient phenomenon of multilingual speech is language mixing. In multilingual communities, it is common for people to switch between languages in different ways. This can occur at very different points within a conversation. For example, a topic change within a conversation may trigger a language switch and the speaker either switches within an utterance or s/he switches the languages between utterance boundaries (traditionally referred to as intra- vs. inter-sentential). For decades, researchers have explored why and how multilinguals switch from one language to another, which cumulated in various theories and models that were proposed (see Adamou and Matras 2020 for an overview of language contact phenomena). One term that is inextricably linked to the description of language mixtures is code-switching, or code-mixing. Code-switching is generally referred to as the switching between languages or varieties within a communicative interaction without hesitations or pauses (Milroy and Muysken 1995, p. 7). The following example of an adult code-switching illustrates a switch from German to English and then back to German.

- (1) Es war Mr Fred Burger, der wohnte da in Gnadenthal and he went out there one day and Mrs Roehr said to him: Wer sind denn die Männer do her? (Clyne 1994, p. 112) 'It was Mr Fred Burger, who lived there in Gnadenthal and he went out there one day and Mrs Roehr said to him: Who are these men?'

As the example illustrates, languages in contact almost certainly influence each other, and the most frequent and most salient contact phenomenon is code-switching. Many disciplines have tried to uncover the underlying reasons for code-switching and have provided various explanations, definitions, and terms. Apart from code-switching, code-mixing, language switching, style shifting, language shifting, or code shifting are other

terms that have been used to describe the use of bi- or multilingual utterances. There is often little agreement about what exactly is captured by these terms and how they differ from each other. The result is that, on the one hand, the same phenomenon is described using different terms and, on the other hand, the same term is used for different phenomena, thus literally a term-switching. This is also stated by Romaine (1995, p. 124) as follows: “In general, in the study of language contact there has been little agreement on the appropriate definitions of various effects of language contact.” Clyne (2003, p. 70) even speaks of a “troublesome terminology around ‘code-switching’”.

According to Muysken (2000, p. 3), the term code-switching has often been used for language switches between clauses, while code-mixing tends to be used for intra-clausal phenomena. As the present paper investigates language switches of a bilingual child within utterances, we follow the terminology also commonly used in language acquisition research by speaking of code-mixing (Di Sciullo et al. 1986; Cantone 2007, pp. 15, 56; Quick and Hartmann 2021). A typical example of child code-mixing is the following:

- (2) was you as well ganz ganz lieb. (Fion, 3;8<sup>1</sup>)  
 ‘were you as well very very good’

The lack of agreement concerning terminology resulted from the different research interests expressed by the various fields investigating the topic. Whereas sociolinguistic approaches were primarily interested in the social functions of code-mixing (Auer and Eastman 2010; Blom and Gumperz 1972; Gumperz and Hymes 1972; Gumperz 1982, among many others), structurally oriented scholars have dealt with code-mixing from a grammatical perspective and the question of possible underlying syntactic constraints. The latter group of linguists were mainly focused on the question of where in a bilingual utterance switching is possible and whether this can be explained by the basic syntactic architecture of languages and their typological features. The concern of these studies on code-mixing is primarily to establish universal grammatical constraints on where and when in a sentence language switching can occur (Belazi et al. 1994; MacSwan 1999; Myers-Scotton 1993; Poplack 1980, among many others). Universal here means that the constraint applies to any language pair. Functional categories were thought to play a key role here. A very influential model that addresses this idea is the so-called matrix language frame model (MLF) by Myers-Scotton (2006). The model assumes that one language, the so-called matrix language, forms the grammatical framework in which words or phrases of another language are embedded (embedded language). The matrix language dominates code-mixed utterances by determining both word and morpheme order (morpheme order principle) and by prescribing all functional elements (system morpheme principle) (Myers-Scotton and Jake 2015, p. 421). In general, this approach has been shown to describe some code-mixing patterns quite well, including studies of code-mixing in children (Paradis et al. 2000; M. Vihman 1998; V. Vihman 2018). However, other studies come to results that contradict the predictions of the approach (Backus 2014; Quick et al. 2018a, 2018b, 2018c, 2021; V. Vihman 2018; Zabrodskaia 2009).

Fundamental criticism also comes from formal language theories. MacSwan (2000, 2005), for example, considers the differentiation into a matrix language and an embedded language to be implausible. Rather, he suggests that code-mixing research should refrain from formulating specific constraints that apply only to code-mixing, such as those assumed in the MLF model (MacSwan 2005, p. 20). Instead, code-mixing should be embedded in a general language theory. As such, he proposes the minimalist program (Chomsky 1995, among others). That is, code-mixing should not be constrained by anything apart from the properties of the grammars involved themselves. Restrictions applying to code-mixed utterances only are rejected.

Thus, it can be stated that Myers-Scotton’s idea of a matrix language and MacSwan’s demand for restrictions that do not explicitly apply to code-mixing but are integrated into a theory of language, i.e., which are also valid for monolingual utterance, are relatively irreconcilably opposed to each other. A promising alternative to the explanation of code-

mixing, in which both points of view are taken into account, is a usage-based approach (Tomasello 2003). In the following, we will present this as a possible alternative explanation for code-mixing, especially for child code-mixing. Additionally, we will discuss how code-mixing can be methodologically approached from a usage-based perspective.

## 2. Usage-Based Approaches in Language Acquisition Theory

One of the main questions that the field of language acquisition has been concerned with revolves around the issue of whether languages are acquired largely based on innate prerequisites of languages, or whether children are able to actively construct their languages from the input they receive (see Ambridge and Lieven 2011 for an overview). The latter view has been gaining popularity, especially in the form of constructivist accounts. The basic idea here is that language can be constructed from the interplay of general social-cognitive abilities and the language(s) offered by the environment. Thus, these approaches follow, in their basic orientation, the idea of epigenesis: behavior, in this case language, and the neural system underlying it, emerge from the interaction of neurophysiological processes and experience-based events (Szagun 2019, p. 265). Accordingly, language cannot be solely related to genes or to the environment. This basic assumption is followed by a number of approaches, some of which focus on different areas of language acquisition (Klann-Delius 2016). In the following, we will focus on the usage-based approach (Tomasello 2003, among others), as it offers very concrete alternative assumptions for the acquisition of grammar. For example, in contrast to nativist approaches, usage-based approaches do not assume that there is an innate blueprint for the acquisition of grammar. This does not necessarily mean that the acquisition of language does not rely on innate cognitive capacities at all. However, instead of assuming language-specific “modules” of cognition, usage-based approaches see language as a mainly cultural phenomenon that builds on and is closely interwoven with domain-general cognitive capacities such as pattern-finding and intention reading (see e.g., Tomasello 2015). This results in a diametrically different view on language acquisition: grammatical structures are not stored a priori as abstract rule-based knowledge that is applied in a top-down process, but language (grammar) must be built up gradually in a bottom-up process. Nativist approaches base their ideas on an innate grammar: language is assumed to be organized in a modular form with a strict separation of lexicon and grammar (grammar = abstract syntactic structures). On the contrary, usage-based approaches are guided by assumptions from cognitive functional linguistics (Bybee 2010, among others) and construction grammar (Goldberg 2006, 2019). They reject the idea of modularity and the strict separation of lexicon and grammar, but instead assume that language consists to a large extent of fixed word combinations and frame-and-slot patterns. Essentially, this assumption blurs the line between lexicon and grammar. Usage-based approaches therefore assume that grammar and lexicon form a continuum. According to this view, language consists not only of words and morphemes and syntactic rules according to which these words and morphemes are combined, but of constructions of varying complexity—this varying degree of complexity also plays a central role in children’s code-mixed utterances (see Section 2.2). Constructions can be divided into three dimensions of linguistic knowledge that are of particular relevance for language acquisition (Koch 2019, p. 60f.):

- I. Specific constructions (chunks): fully lexicalized constructions such as (complex) words, fixed multi-word utterances, grammatical phrasemes, and idioms.
- II. Frame-and-slot patterns: partially lexicalized constructions such as derivational and inflectional morphemes (VERB-ed, e.g., *play-ed*, *seek-ed*, *smil-ed*), schematic idioms (*the X-er the Y-er*), or schematic constructions with lexically specific elements and empty slots (*I see X*; *Give me X*).
- III. Abstract constructions<sup>2</sup>: fully schematic constructions [Subj V Obj1 Obj2].

This means language can vary in the level of complexity, ranging from individual words to large units of processing. Children are just language beginners and their emerging language is characterized by a large amount of unit-like constructions. Contrary to nativist

accounts, usage-based approaches refuse an initial abstractness of children's grammar. Instead, they emphasize the notion of a lexically specific, schematic, and fully abstract construction-like taxonomy. At the beginning of the language acquisition process, children acquire specific knowledge and arrive at abstract schemata via partially abstract schemata, with unanalyzed frame elements (fixed lexical units) and empty slots. The prerequisites for this are presented in the following section.

### 2.1. The Role of Input and Pattern Finding

Nativist approaches base their idea of a genetic embedding of language on a presumed poverty of the linguistic input. Thus, the linguistic environment to which children are exposed is insufficient for them to acquire the key structural principles of a language within a relatively short period of time: "How do we come to have such rich and specific knowledge, or such intricate systems of belief and understanding, when the evidence available to us is so meager?" (Chomsky 1987, p. 33). The asymmetry of an impoverished input on the one hand and the complexity of linguistic systems on the other led Chomsky to conclude that there is an explanatory problem, namely: "The problem of poverty of stimulus." (Chomsky 1986, p. 7). As a consequence, he proposed that only with the help of an innate language knowledge are children able to acquire a finite set of rules, which in turn enables them to use language productively and creatively (Koch 2019, pp. 10–12).

Numerous studies have challenged the innateness assumption and presented evidence in favor of alternative cognitively oriented approaches (e.g., MacWhinney 2004; Tomasello 2003). Input, especially child-directed speech (CDS), is by no means as ungrammatical and impoverished as Chomsky and colleagues have claimed. For example, in their study of several corpora of parent–child interactions, Sagae et al. (2004) showed that the input is comparable in quality to the corpora of the *Wall Street Journal*. A study by Cameron-Faulkner et al. (2003) has also shown that CDS is not chaotic at all but exhibits a high degree of lexical and grammatical repetitiveness (2003). Analyzing the language use of 12 English-speaking mothers, the authors could show that the mothers were extremely repetitive in that they had a high number of lexically constrained utterance-initial patterns, such as *Where is X? What is X?* and *Are you X?* At the same time, there was a high correlation between these item-based linguistic patterns of the mothers with the utterances of their 2-year-old children. One could argue that the results are not surprising because of the typological properties of English with its relatively fixed word order. However, Stoll et al. (2009) analyzed mothers' CDS in Russian and German, languages that have a relatively variable word order, and found again that the input showed a high degree of repetitiveness in the initial sentence patterns.

The correlation of linguistic input and output is also shown in a study on a German child, Leo (Behrens 2006). This study also provided further evidence that children are able to extract linguistic knowledge from the input they receive, thus challenging the idea of an impoverished input. Behrens (2006) focused on the question of how children's output is influenced by the input they receive from their immediate caregivers. For this purpose, the distribution of word types and the internal constituent organization of nominal and verbal phrases was analyzed over a period of three years. While the adult input data show a homogeneous distribution over the period under consideration, the child output steadily approaches this distribution over time. Toward the end of the study, at the age of 4;4 to 4;11, Leo has almost completely adapted to the distribution pattern of the input and thus to his caregivers. Furthermore, the study shows that children hear several thousand utterances from their caregivers every day, in some cases more than 3500 words per hour. For this reason, the role of input plays a vital role in the study of language acquisition from a usage-based perspective (see e.g., Rowland et al. 2003).

According to this view, then, children receive rich linguistic input, which enables them to discover regularities and thus to build up their grammars. Although language is extremely varied, it is also very structured, and children acquire their native languages simply by tracking patterns in the input and generalizing from them, but how exactly

does this work? From a usage-based perspective it is assumed that the acquisition of linguistic structures is based on general cognitive processes. Here, the domain-general ability of humans to recognize patterns (pattern-finding) is considered to be central for language acquisition (Tomasello 2003, p. 34). From early on, children are able to filter out and categorize regularities of various kinds from language examples of the input. Accordingly, the rules of a language are continuously built up from language use in a generalization process. The communicative contexts prevailing at the beginning of language acquisition facilitate this for children, since the language addressed to them is highly repetitive, containing numerous repetitions and a high number of formulaic utterances (Szagun 2019, pp. 206–38). In addition, already recognized patterns help children to further perceive and categorize their linguistic environment (Romberg and Saffran 2010). This paper focuses on the methodological possibilities for identifying such linguistic patterns in corpora. Two of these current methods are presented in more detail below and then applied to the linguistic data of a child. Before that, however, a brief look at the development and construction of linguistic patterns, especially multilingual patterns, will be taken.

## 2.2. Building Up Grammatical Knowledge: From Concrete Utterances to Abstract Schemas and Their Role in Code-Mixing

From a usage-based perspective, children acquire their language(s) based on their social-cognitive skills. Children's earliest words and utterances are often direct imitations of the input. These can be single words, but also fixed word combinations. These unanalyzed formulaic set pieces are holistically stored as units, so-called chunks. Examples for chunks include multi-word utterances like *all\_away* or *what\_is\_that?*, which can be used by children without recognizing that these are multiple interconnected words. For this reason, MacWhinney (2014, p. 34) argues that the transition between children's first words and their first sentences is barely perceptible. From a usage-based perspective, children are thus initially primarily conservative learners who imitate utterances of their caregivers (e.g., Bannard and Matthews 2008; Bannard et al. 2009; Tomasello 2003). However, the process of language acquisition is not limited to imitative learning solely. As mentioned earlier, children's pattern recognition skills help them detect similarities and filter out recurring elements in the input. In this way, they begin to form linguistic categories and thus acquire a large amount of abstract knowledge based on the linguistic constructions they encounter. This happens in an item-based manner, which means that linguistic knowledge is built up around concrete words and phrases. Similar utterances such as *I want ice cream!*, *I want Fiene-Biene!*, and *I want that!*, which are linked to recurring patterns of perception or action, are combined in a process of abstraction to form the schema *I want X!* (Koch 2019, p. 139). This contains the empty slot (*X*) in addition to the fixed lexical elements (*I* and *want*). Children gradually start to use the construction productively by filling the open slot with different words to ask for further specific objects (*I want chocolate!*, *I want Mommy!*). That is, as children move from lexically specific units to more schematic constructions, they dissect concrete utterances and identify the gaps that are there and that can be filled productively. However, even in these constructions, children are only productive in limited ways and in a piecemeal way gradually create abstract categories and schematic constructions (e.g., Lieven et al. 2009; Tomasello and Brooks 1999). Pattern finding enables the children to go beyond item-based constructions and to form abstract categories and schemas. Consequently, children acquire their linguistic knowledge in a piecemeal way, starting out with fixed chunks (holophrases) such as *gimme this* and frame-and-slot patterns like *I want X*. The item-based nature in children's early speech has been shown in numerous corpus-based and experimental studies for different languages (e.g., Bannard and Matthews 2008; Lieven et al. 2009).

So far, however, usage-based studies have largely focused on monolingual acquisition. In a global perspective, though, monolingual language acquisition is the exception rather than the norm (see Crystal 2003, p. 17). A key topic in language acquisition research,

therefore, is the question of how multilingual children acquire their languages and how these languages influence each other. One particularly salient contact phenomenon in multilingual language acquisition is code-mixing. Code-mixing has been discussed extensively in the literature on multilingual language acquisition (e.g., [Cantone 2007](#), among many others). However, one aspect that has arguably been neglected so far is the role of input and its influence on a child's code-mixed patterns. As mentioned before, work on language acquisition has shown that there is an intimate relationship between linguistic knowledge and the instances of language use: input provides the basis from which children extract and build up their linguistic knowledge. In particular, children abstract away fixed chunks like *what's this?* as well as frame-and-slot patterns like *what's X?* from the input they receive, and these patterns serve as the basis for the children's own productive language use. From a usage-based perspective, it can reasonably be assumed that these constructional patterns also form the basis for code-mixed utterances, e.g., *I got X* as in *I got noch eine* 'I got another one' (Lily, 2;4). To detect those patterns, one needs reliable methods. In addition to corpus and experimental evidence, several studies have taken a computational approach (e.g., [Bannard et al. 2009](#); [Lieven et al. 2009](#); [McCauley and Christiansen 2017](#)). In the following, two of these methods, the traceback method (TB) and the chunk-based learner (CBL), are examined in more detail and applied to a child's linguistic data, and, at the end, their advantages and disadvantages will be discussed. Both models were originally developed to account for monolingual or L2 data, but recent studies have applied the TB to code-mixing, showing that item-specificity also plays a role in these utterances. We want to verify this and at the same time compare another method, the CBL, with the TB method applied to code-mixed data.

### 3. Corpus Study

#### 3.1. Participant

For the present study, we investigate an English–German bilingual boy, Fion. The boy grew up in a one-parent-one-language (OPOL) household with his mother being a native speaker of German and his father being a native speaker of English. Although Fion's parents mostly adhered to the OPOL strategy, they did not commit to one family language and sometimes used both languages alternately when everyone was present. Furthermore, Fion has an older brother who also grew up bilingually from birth and sometimes used code-mixing when talking to Fion or his parents. The family lived in Germany and belonged to a middle-class household. At 18 months, Fion attended a German-speaking daycare 5 days a week.

#### 3.2. Data

The analysis is based on Fion's spontaneous speech recordings, which were collected during the age of 2;3–3;11. On average, two hours per week were recorded in the family's home, resulting in a total of 205 h for the entire recording period. In total, 47,812 child utterances (3492 of which included code-mixing) and 180,293 input utterances were included in the analysis. The data were transcribed in CHAT format ([MacWhinney 2000](#)) and enriched with a small set of annotations by Antje Endesfelder Quick.

#### 3.3. Methods

As mentioned above, our study is couched in the framework of a usage-based approach, which sees speakers' grammars as fundamentally grounded in concrete usage events. On the methodological side, this translates into a thoroughly data-oriented approach that has yielded highly relevant insights, especially in research on monolingual language acquisition (see e.g., [Behrens 2009](#); [Tomasello 2015](#) for overviews). We propose that the same methods can be fruitfully applied to multilingual acquisition, focusing on the relationship between input and pattern realization. In particular, we study the use of code-mixing, which can give valuable clues to how constructional patterns from different languages interact. In the present study, we compare the application of two methods.

First, data are analyzed using the TB method. For this purpose, we briefly summarize a study conducted by [Quick and Hartmann \(2021\)](#), in which a variant of the method was used. In the following, we first briefly introduce the basic idea of TB and then explain its adaptation for the analysis of Fion's code-mixed data. After that, the main features of the second method, the chunk-based learner (CBL), are explained, as well as its application to bilingual data.

### 3.3.1. The Traceback Method

Numerous corpus-based studies of language acquisition have relied on different variants of TB. Its basic idea was already anticipated in [Lieven et al. \(1997\)](#) and first elaborated systematically in [Lieven et al. \(2003\)](#) and finally specified in [Dąbrowska and Lieven \(2005\)](#). In a series of previous studies, we have used the TB method (e.g., [Quick et al. 2018a, 2018b, 2018c, 2021; Quick and Hartmann 2021](#)). As shown in [Hartmann et al. \(2021\)](#), the TB method has originally been developed for testing the hypothesis that much of early child language can be accounted for with the help of a fairly limited set of patterns. Later on, the method was also used in a more explorative way to investigate those patterns in more detail. The basic idea of TB is simple: a longitudinal child language corpus is split in two parts, a main corpus and a test corpus, the latter usually comprising the last one or two session(s) of recording. The child's utterances in the test corpus, the so-called target utterances, are then "traced back" to previous utterances in the main corpus. This means that, first of all, the main corpus is searched for verbatim matches of the target utterance. For example, if the target utterance is *I want this one*, the algorithm (or, in some implementations, a human analyst) checks whether the exact same utterance occurs in the main corpus. Usually, both the child's utterances and the caregivers' input are taken into account here. Thus, if the utterance occurs in the main corpus, the traceback is considered successful for this particular utterance, no matter if the attestation in the main corpus occurs in the child's or in the caregivers' data. If there is a verbatim match in the main corpus, we speak of "fixed strings" or "fixed chunks". In many cases, there is of course no verbatim match. In this case, the main corpus is queried for partial matches in a principled way, using a set of so-called operations that are used to "construct" the target utterance from "component units" that can be found in the main corpus (and each of which has to be attested at least twice in most implementations of the TB method). For example, if the target utterance is *I want a paper*, and there is no verbatim match in the main corpus, it is still possible that the child has a so-called frame-and-slot pattern such as [I want THING]<sup>3</sup> or [I want a THING] available, consisting of a fixed part and a variable slot. If the main corpus contains utterances like *I want a toy* and *I want a cake*, we can posit the frame-and-slot pattern [I want a THING]. If *paper* is attested at least twice in the main corpus as well, *I want a paper* can be accounted for as being "constructed" of the component units [I want a THING] and *paper*. To "construct" the utterance, the operation SUBSTITUTE is used, one of three operations that are used in most TB studies, the other two being SUPERIMPOSE and ADD (some TB studies use additional operations, e.g., [Lieven et al. 2003](#)). (3) illustrates each of the three operations. SUPERIMPOSE and SUBSTITUTE are very similar to each other, but while in the case of SUBSTITUTE a component unit is inserted in the open slot of a schema (e.g., *teddy* fills the REF(erent) slot in [I want my REF] in (3)), there is some overlap between the fixed string within the schema (e.g., [I want my REF] in (3)) and the component unit (e.g., *my teddy*). ADD does not involve schemas with an open slot but rather refers to the juxtaposition of fixed strings. To prevent the method from being too unconstrained, ADD is usually restricted to items that can be positioned before or after the adjacent component unit, e.g., *I want my teddy now* or *Now I want my teddy* in (3).

(3)	SUBSTITUTE	I	want	my	REF	teddy	→	I want my teddy.
	SUPERIMPOSE	I	want	my	REF	teddy	→	I want my teddy.
	ADD	I	want	my	teddy		→	I want my teddy now.
			+ now					or: Now I want my teddy.

Depending on how the authors operationalize the TB method, it can yield different results. For example, results can differ based on which operations one assumes and which threshold is set for a traceback to count as successful (some of the parameters along which different implementations of the method vary are mentioned in Figure 1 below; also see Koch et al. 2020). Still, a number of studies have shown that regardless of the exact implementation, a large proportion of target utterances can be successfully traced back (see e.g., Dąbrowska 2014; Dąbrowska and Lieven 2005; Vogt and Lieven 2010; for Italian, see Miorelli 2017; for German, see Koch 2019; for critical evaluations of the method, see Koch et al. 2020; Kol et al. 2014). This provides evidence in favor of the usage-based hypothesis that language acquisition is strongly item-based and that children are able to actively construct the complexity of their language(s).

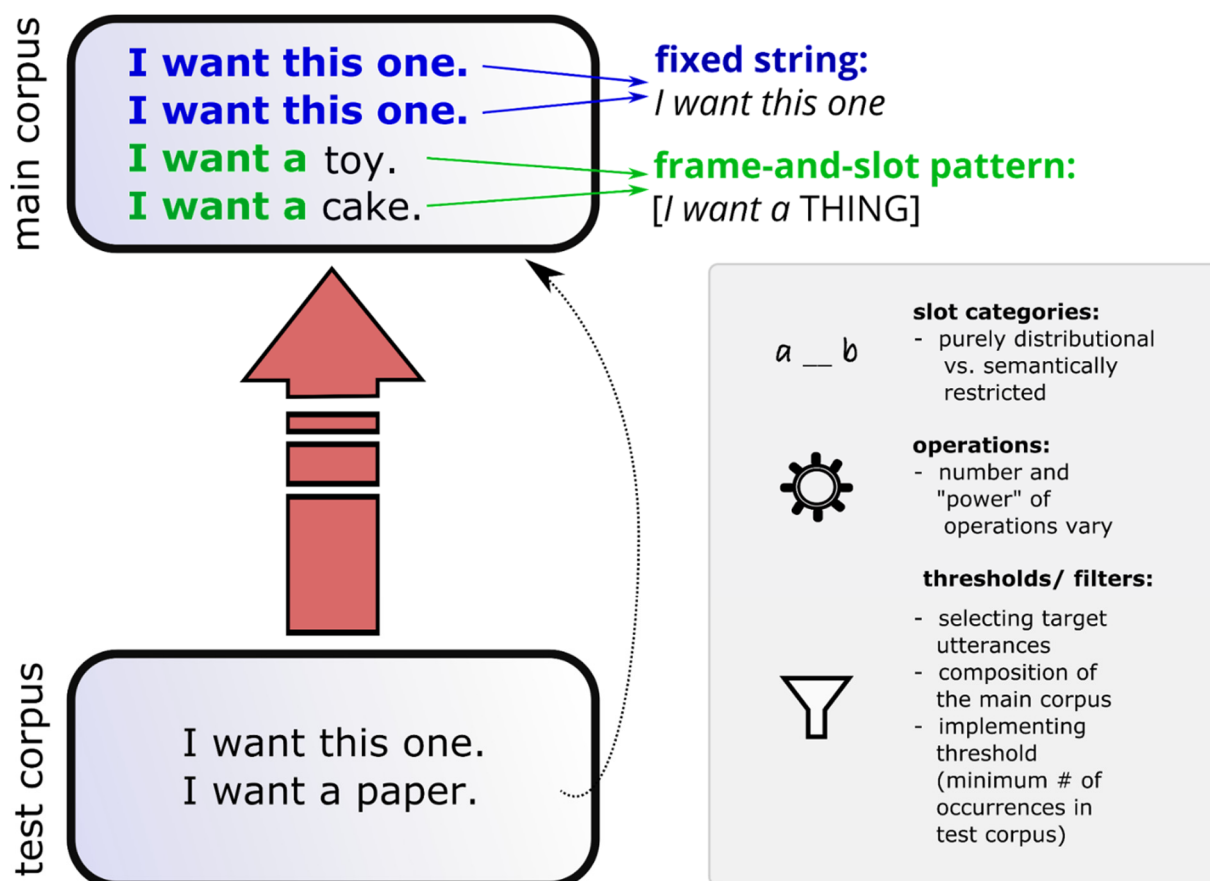


Figure 1. The TB method, illustrated with (constructed) examples from Dąbrowska and Lieven (2005). The grey box shows the “parameters” along which different implementations of the TB method vary. Figure from Hartmann et al. (2021), CC-BY 4.0.

In Quick and Hartmann (2021), the Fion data are analyzed in a slightly different way using a variant of TB. Rather than using the last session(s) of the recording as the test corpus,



they use the code-mixed data as the test corpus and trace the multilingual utterances back to both the monolingual ones and to the caregivers' input. In addition, the full set of child utterances is traced to the input data. Thus, the application of the TB method used here boils down to "cross-tracing" rather than "traceback", as the target utterances are not traced to earlier utterances but rather to utterances of a different type. (We will still stick to the term "traceback" here as it is the established name for the method.) The reasoning behind this operationalization is that if a child's utterances can be successfully accounted for with a limited set of patterns—as has been shown repeatedly for monolingual children in previous TB studies—and if one further assumes that children learn these patterns from their linguistic experience, then one should be able to trace the child's utterances, even the code-mixed ones, to the input, just as one can trace them back to the child's own previous utterances. In this way, [Quick and Hartmann \(2021\)](#) checked to what extent the child's code-mixed utterances can be accounted for with the help of frame-and-slot patterns and fixed chunks that Fion either used himself or that he has heard in the input.

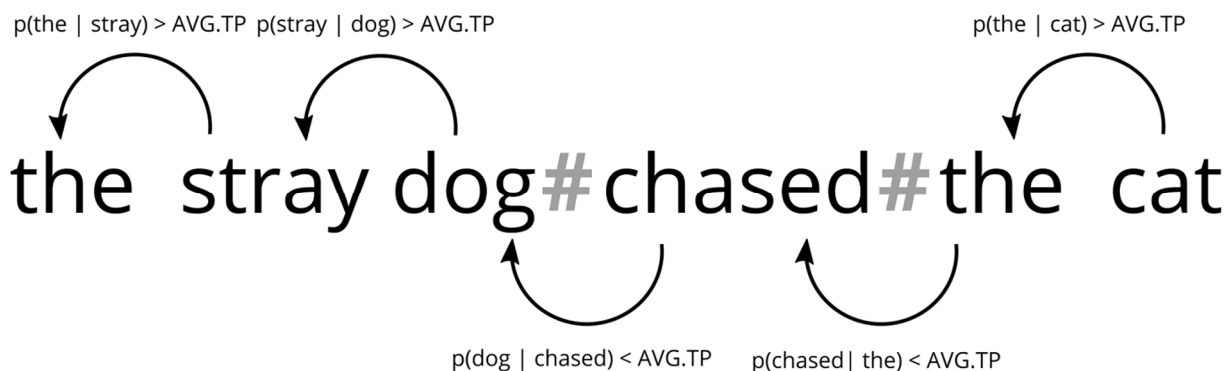
### 3.3.2. Chunk-Based Learner (CBL)

The second methodological approach we draw on is the chunk-based learner (CBL). The CBL combines simple frequency measures with psychologically inspired learning mechanisms to develop a (predictive) grammar of a language learner based on authentic data. This method has been used to carve out commonalities and differences between first and second language learning ([McCauley and Christiansen 2017](#)) and can contribute significantly to our understanding of bilingual language learning. Similar to the TB method, the corpus is split in two parts: about 90% of the corpus is used as training data for the computational model (comprehension phase), which can then be used to evaluate the model's ability to produce utterances like the ones in the portion of the corpus that was withheld from the model (production phase). In the production phase, the model receives the utterances from the withheld part of the corpus as an unordered bag of words and then attempts to reconstruct the original utterance based on its training data. [McCauley and Christiansen \(2019\)](#) have used this model to compare first- and second-language learners, showing that raw frequency-based chunks provide a better fit to L2 learning data, while sequence coherence seems to play a bigger role in L1 acquisition.

The CBL method was originally developed by [McCauley and Christiansen \(2011\)](#) under the name "CAPPUCCINO model". Later on, it was refined and applied to different kinds of data, including L2 learner data ([McCauley et al. 2015](#); [McCauley and Christiansen 2014, 2017, 2019](#)). CAPPUCCINO stands for "comprehension and production performed using chunks computed incrementally, non-categorically, and on-line", which already reveals much about the way the model works. The basic idea is to incrementally detect chunks using relatively simple statistics. In other words, the model "learns" linguistic units by being fed with input word by word and then detecting structure in that input in a bottom-up way. To recognize multi-word units that belong together, the model uses a simple metric: backward transitional probabilities (BTP). One of the reasons why this metric was chosen is that a series of studies have shown that even very young children are sensitive to BTPs (see [Pelucchi et al. 2009](#)). The BTP value indicates how likely it is that the current word is preceded by the word that precedes it in the current context. The BTP value is calculated exclusively on the basis of the units processed so far. For example, if the algorithm first encounters the utterance *What is this?* as illustrated in 4 (a) and then the utterance *What was this?* (4b), the BTP for all word pairs in the first utterance will be 1, as the model only knows this first utterance at this point in time, as shown in (4a). (The  $\hat{\ }^{\wedge}$  sign denotes the beginning of the utterance, which is treated as a token in its own right.) As soon as the model encounters the second utterance, the likelihood that *this* is preceded by *is* (or *was*, for that matter) decreases to 0.5 as the model has now seen two different words preceding it.

(4)	a. <sup>^</sup>	←	what	←	is	←	this
			btp		1		1
			avg.btp		(1)		(1)
	b. <sup>^</sup>	←	what	←	was	←	this
			Btp		1		0.5
			avg.btp		(1)		(0.92)

Apart from the BTP value, which is calculated for each word pair on the basis of the utterances the model has encountered so far, the algorithm also calculates the average BTP value of all word pairs that have entered the model so far, shown below the BTP values in Figure 2. To identify multi-word units that belong together, the model proceeds as follows: if the current BTP value is above the average BTP, the word pair is regarded as a potential chunk. If the BTP of the word pair falls below the current average BTP, a boundary between the two words is assumed, and the first word is grouped together with the preceding word(s), as Figure 2 demonstrates, using the example of *the stray dog chased the cat*. Assuming that the BTPs between *the* and *stray*, as well as between *stray* and *dog*, are above the average BTP in each case, while the probability of *chased* being preceded by *dog* falls below the average BTP, the model assumes a boundary between *dog* and *chased* (marked by # in the figure). Thus, *the stray dog* is regarded as a chunk. As the first split is inserted after *dog*, *dog* is grouped with the preceding word pair. Chunks can of course also consist of more than three words, as a chunk always contains all material between two boundaries (#). Apart from the splits that are calculated by comparing BTP and average BTP, the beginning and end of an utterance are boundaries as well.



**Figure 2.** Simplified illustration of the chunk-based learner algorithm, freely adapted from [McCauley and Christiansen \(2019, p. 10\)](#). The # signs show where the algorithm posits a boundary between chunks because, in those positions, the backward transitional probability between the lexemes  $p(\text{lexeme1} | \text{lexeme2})$  falls below the current average transitional probability (avg.tp).

As the multi-word unit *the stray dog* has now been identified as a chunk, it enters the inventory of chunks, the so-called chunkatory. This chunkatory is also used by the model for processing the utterances it encounters: Each word pair that enters the model will first be searched for in the chunkatory, and if it is already present there—either as a chunk in its own right or as part of a larger chunk—the two words will automatically be treated as belonging together, and no boundary will be drawn between them, as long as the chunk is attested at least twice in the data ([McCauley and Christiansen 2019, p. 9](#)).

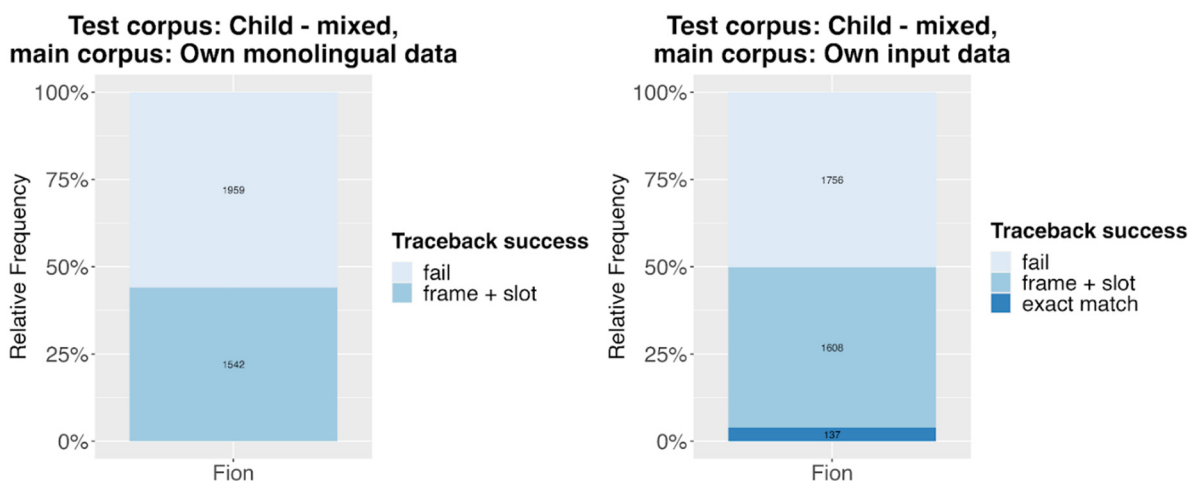
As already pointed out at the beginning of this section, the CBL model consists of two components, which are intended to model language production and comprehension in a simplified way. The steps described so far all belong to the “comprehension” part of the model, which is also what we will focus on for the purposes of the present paper. McCauley and Christiansen’s implementation of the CBL model goes further, however, by using the chunks that the model has identified for predicting the utterances of the child. We refer to [McCauley and Christiansen \(2019, p. 14f.\)](#) for more details; in the present paper,

only the comprehension side of the model will be taken into account, as we are interested in the chunks identified by the CBL algorithm.

#### 4. Results

##### 4.1. Code-Mixing and the Traceback Method

The TB results of Fion’s code-mixed utterances are summarized in Figure 3. All code-mixed utterances were initially traced back to his own monolingual data (left graph) and, in a second step, also to the caregivers’ input (right graph). First, we can note that the TB success rate is lower compared to the other TB studies. TB studies investigating German monolingual language acquisition presented successful TB scores ranging from 85% to 95% for children between 2;0 and 2;6 years. It has to be noted that 6% to 24% verbatim matches from the main corpus were included. This is different for Fion (see right graph) for several reasons. First, verbatim matches with the code-mixed utterances occur much less frequently as, on the one hand, code-mixed utterances are a low frequency phenomenon and, second, the test corpora contained only code-mixed utterances, whereas the main corpora consisted almost exclusively of monolingual utterances. The only exceptions here are the very few code-mixed utterances in the input (especially from Fion’s brother). Fion’s own monolingual data, of course, contain only monolingual utterances. Overall, almost 50% of the code-mixed utterances can be successfully traced back when all input is considered and slightly less when only Fion’s monolingual utterances are considered.



**Figure 3.** TB results. Fion’s code-mixed data traced back to his monolingual data and to the input. Figure adapted from Quick and Hartmann (2021).

Given that the code-mixed data were used as target utterances, these numbers are higher than one might expect. The majority of the code-mixed utterances are partial tracebacks and render frame-and-slot patterns. This is quite remarkable since Fion’s input data is exclusively monolingual.

Thus, the results suggest that Fion has patterns available in his monolingual data and/or in the input from which he can extract patterns to build his code-mixed utterances as well. Table 1 shows the most frequently attested patterns for Fion. The following utterances (5) illustrate this:

- (5) a. und jetzt mach this. (Fion, 2;3)  
'and now make/do this'
- b. noch mehr grapes. (Fion, 2;5)  
'more grapes'
- c. ich will one more. (Fion, 2;4)  
'I want one more'

**Table 1.** Most frequently attested patterns in the Fion data according to the TB method, from [Quick and Hartmann \(2021\)](#).

Pattern	Frequency Fion
__ this	100
__ da	20
noch mehr __	16
__ that	15
ein __	11
ich will __	11
__ you	11
und ein __	10
__ cheese	9
__ feuer	9
__ nicht	9
nein ein __	8
__ milk	8
das istis __	7
ich __	7
this __	7
was ist __	7
wo eine __	7
wo istis __	7
wo kommt __	7

All three can be explained with frame-and-slot patterns (see Table 1), which Fion also uses in his monolingual utterances. It is important to note that, in all cases, not only the frame-and-slot patterns are attested in the main corpus but also the slot fillers that occur in the open slots in the individual utterances. It is, however, a matter of debate to what extent the patterns identified by TB are cognitively plausible (see e.g., [Hartmann et al. 2021](#)). In our operationalization, first and foremost, they serve as a proof of concept that the code-mixed utterances can, in principle, be accounted for using frame-and-slot patterns.

The results only show that the patterns found in the code-mixed utterances can be traced back to the input Fion received, suggesting that children are able to pick up linguistic knowledge from their environment. This can be seen as evidence in favor of the hypothesis that the usage-based model of language acquisition can also account for code-mixed data, as language-mixed utterances can be analyzed as instances of either fixed chunks or frame-and-slot patterns. However, as pointed out in various TB studies before, as well as in methodologically oriented papers (see e.g., [Hartmann et al. 2021](#); [Koch et al. 2020](#)), the method has a number of limitations. Perhaps most importantly, not all of the patterns identified via TB can be assumed to be psychologically plausible. This is why it makes sense to combine TB with other methods that take a different approach to pattern identification. One of them is the chunk-based learner (CBL), to which we now turn.

#### 4.2. Code-Mixing Meets the CBL Model

In light of the discussion above, it can prove insightful to take a closer look at the results of the CBL model when it encounters code-mixed data. As mentioned above, we will only focus on the “comprehension” side here, i.e., the chunks that the model identifies, without evaluating the results on the basis of out-of-bag data using the production side of the model.

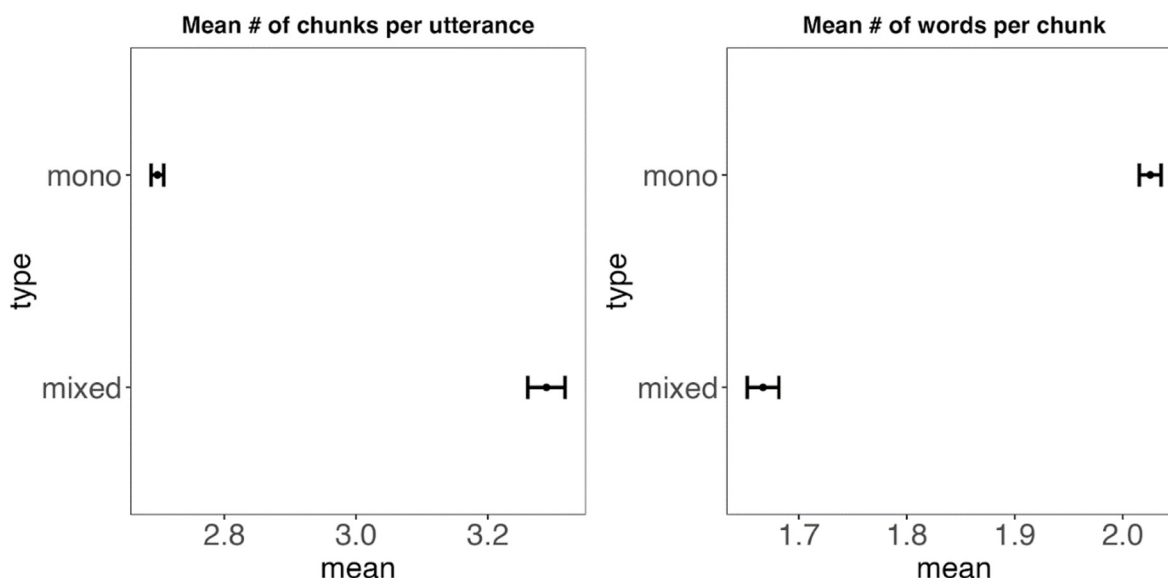
We used McCauley’s CBL script, available at <https://github.com/StewartMcCauley/CBL/> (accessed on 5 September 2022), to obtain the chunking results. The entire Fion dataset, including Fion’s monolingual utterances and the caregivers’ speech, was used as input for the CBL algorithm.

As in the case of the TB results, we analyze the results both quantitatively and qualitatively. As for the quantitative perspective, it can be instructive to look at differences between the chunks identified in the child’s code-mixed and his monolingual utterances. When processing the input it receives, the model does not make a difference between them because it cannot know which of the utterances it parses are German, which are English, and which are code-mixed. Unsurprisingly, the number of chunks per utterance that the model identifies tends to be lower in monolingual utterances than in code-mixed ones, while the mean number of words per chunk tends to be higher in monolingual utterances, as shown in Figure 4. This result is expected because it stands to reason that the boundary chunk is often placed at the position of the language switch (because an English word is, of course, relatively unlikely to be preceded by a German one, and vice versa), as in the following examples:

- (6) a. ein kleinen || shark (Fion, 2;3.16)
- b. nein || a || nein || a ice hockey player (Fion, 2;3.16)
- c. zeig || ice cream (Fion, 2;3.16)

Indeed, the position of chunk boundaries identified by the CBL algorithm often coincides with the position of code-switches. In 1120 cases, all code-switches in the respective utterance coincide with chunk boundaries; in 1960 cases, there is at least a partial overlap between the position of chunk boundaries and of language switches. Only in 426 code-mixed utterances, there is no overlap. In those utterances, we often find recurrent code-mixed sequences such as *time out machen* ‘do time out’ or *(ein) anderer frog* ‘another frog’, each of which is attested five times in the corpus.

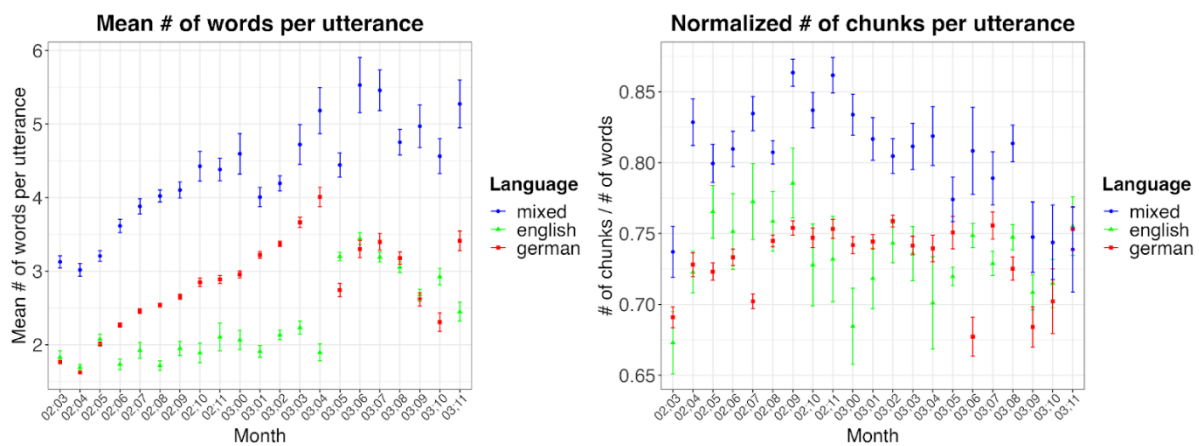
As there are no language switches in the monolingual data, the likelihood is higher that longer sequences of words are recognized as chunks, compared to the code-switched ones. This explains why the average number of words per chunk is higher in the case of the monolingual data.



**Figure 4.** Mean number of chunks per utterance and mean number of words per chunk in the code-mixed vs. monolingual data. The error bars show the standard error.

This raises the question of whether we can detect changes between the code-mixed and the monolingual data regarding the number of chunks per utterance. The plot on the

righthand side of Figure 5 shows the number of chunks per utterance. As the number of words per utterance differs, of course, the number of chunks has been normalized by dividing it by the number of words in each utterance. Again, we see that the number of chunks per utterance is much higher in the code-mixed than in the monolingual data. However, the normalized number of chunks in the code-mixed utterances decreases over time. This can be attributed to the gradual limitation of code-mixed utterances to a number of more or less fixed patterns that Fion uses quite frequently.



**Figure 5.** Mean number of words per utterance (left) and the normalized number of chunks per utterance.

Let us now take a more qualitative look at the chunks that have been identified by the CBL algorithm, focusing on chunks comprising more than one word. Table 2 shows the 10 most frequent chunks with 2 or more words and the 10 most frequent chunks with 3 or more words in the monolingual and in the code-mixed data.

**Table 2.** The top 10 most frequent chunks with 2 or more (left) or 10 or more words (right) identified by the CBL algorithm in the child’s code-mixed (upper 10 rows) and the monolingual utterances (lower 10 rows).

Code-Mixed 2+ Words	Frequency	Code-Mixed 3+ Words	Frequency
ich will	63	ich will nicht	14
ein sword	36	choo choo train	5
this is/ist	36	I want to	5
I’m	33	that’s a	5
that’s	33	eine sexy lady	4
ich bin	28	I can’t	4
das ist/is	27	noch mehr orange juice	4
und dann	23	und hier/here is/ist	4
guck mal	22	a builder handy	3
das ist	21	cup of tea	3
Monolingual 2+ Words	Frequency	Monolingual 3+ Words	Frequency
guck mal	723	I can’t	71
ich will	436	ich kann nicht	66
ich bin	428	I want to	49
I’m	265	ich weiss nicht	48
das ist	244	don’t know	47
ich habe	196	I don’t know	40
look at	194	was ist das	39
noch mehr	188	ich will nicht	36
du bist	140	it’s not	33
was denn	138	in the garden	28

Especially in comparison to the TB results, one interesting observation is that very frequent code-mixed sequences like *nein this*, which were identified as fixed chunks by the TB method, are not identified as chunks by the CBL algorithm. This can be seen as a clear advantage of CBL: while TB focuses exclusively on the frequency of the (potential) component units, CBL can take context factors into account much more thoroughly because of the backward transitional probabilities. This leads to the identification of chunks that can arguably be considered cognitively more plausible than the chunks identified by the TB method. While *nein this* does co-occur quite often in the data, as in (7 a–c), only in 65 out of 1447 cases, is *this* preceded by *nein* (and conversely, only in 65 out of 3778 cases, is *nein* followed by *this*). Some of those many other occurrences in which *nein* is followed by another word than *this* or in which *this* is preceded by a word different from *nein* are exemplified in (7 d–e) and (7 e–g), respectively.

- (7)
- |    |  |
|----|--|
| a. | nein this schnee. (Fion, 2;1.13)             |
|    | ‘no this snow’                               |
| b. | nein this geht aber nicht. (Fion, 2;1.15)    |
|    | ‘no but this does not work’                  |
| c. | nein this is keine hose.                     |
|    | ‘no these are not pants’ (Fion, 2;11)        |
| d. | nein leave it. (Fion, 2;1.27)                |
|    | ‘no leave it’                                |
| e. | nein der kriegt this. (Fion, 2;11.3)         |
|    | ‘no he gets this’                            |
| f. | ich will this. (Fion, 02;6.27)               |
|    | ‘I want this’                                |
| g. | kann this auch mitbringen ja? (Fion, 2;6.28) |
|    | ‘I can also bring this, okay?’               |

As Table 2 shows, some of the chunks identified by CBL are actually multilingual, e.g., *ein sword* ‘a sword’, but here it is an entire NP that is identified as a chunk. In the case of *nein this*, by contrast, it is plausible to assume that the two words are independent from one another and the main reason for its identification as a fixed chunk by the TB method is that two highly frequent words happen to co-occur in adjacent position.

Overall, then, the CBL results complement the TB results in a useful way, and it would be interesting to combine the strengths of both methods in future work: TB has the advantage that it uses frame-and-slot patterns rather than just fixed chunks, which makes it possible to take constructions at different levels of abstraction into account; CBL takes context factors into account, thus allowing for the identification of arguably more plausible patterns.

## 5. Discussion and Conclusions

Over the last decades, a number of methods have been developed with the aim of identifying constructional patterns in authentic speech. Most importantly, no method has been proven to be exhaustive. One reason for this is that different methods implement the notion of patterns in different ways. Whereas the TB method focuses on identifying lexically fixed patterns (chunks), as well as frame-and-slot patterns based on (partial) verbatim matches of previously occurring utterances in the data, the CBL processes utterances word by word to reflect the incremental nature of human sentence processing and thereby focuses on chunking processes. Our results suggest that the two methods complement each other in interesting ways. While CBL excels at detecting recurrent chunks, TB helps to automatically identify partially filled constructions (frame-and-slot patterns), which, however, vary considerably in their plausibility.

If we want to identify patterns that can be considered cognitively plausible, we have to take a number of different and partly conflicting aspects into account. As an example, consider the notion of entrenchment, which has recently been thoroughly fleshed out in the work of Schmid (2017, 2020). Roughly speaking, entrenchment refers to the strength of the mental representation of a unit. The goal of pattern identification from a usage-based

point of view is, ideally, to identify units that are entrenched in the speaker's mind, rather than just combinations of units that co-occur by chance. Entrenchment, however, can be conceived of either as a binary concept (a unit is either entrenched or not) or as a gradual one (a unit can be entrenched to different degrees). These perspectives do not exclude each other—a unit can be entrenched to different degrees (gradual view) as soon as it is entrenched in the first place (binary view), and both perspectives can be heuristically valuable depending on the research question at hand. Simplifying matters a bit, we could say that while the TB method operationalizes entrenchment in a binary way, CBL takes frequency effects into account. As such, we argue that the two methods complement each other in identifying constructional patterns as they focus on different aspects of patternhood. While CBL has the advantage that it posits fixed chunks that are arguably more plausible than those identified by TB, as it takes transitional probabilities between word pairs into account, TB has the advantage that it works with frame-and-slot patterns, which allows for positing patterns with variable slots, rather than just fixed chunks.

When it comes to analyzing code-mixing, both patterns arguably lead to important insights. First, the results of both TB and CBL lend support to the idea that we are mostly dealing with instances of “insertional” code-mixing in our data, i.e., units from one language are inserted into patterns—or “frames”—from another language. Second, the results lend support to the usage-based model of language acquisition and show that it can plausibly be applied to code-mixed data. As mentioned above, it would be interesting to extend the CBL model to take frame-and-slot patterns into account, thus combining the best of both worlds. As discussed above, both methods also have limitations; at the same time, we have not fully exploited the potential of both methods in the present paper. For example, we have not investigated the failed tracebacks, which is a common procedure in TB studies and can yield important additional insights, and, as mentioned above, we have refrained from assessing the production performance of CBL as we were mainly interested in the explorative detection of recurrent patterns at this stage. Both aspects should be addressed in follow-up studies.

Both methods, TB and CBL, are based on the assumption that new utterances of children are based on basic lexical patterns, either in the form of frame-and-slot patterns or chunks. Although it would be possible to trace all utterances back to completely abstract structures or rules, as is the case in generative approaches, there is no reason to do so (Lieven et al. 2009, p. 502). The usage-based methods have been able to show that a productive and creative use of language can be explained with frame-and-slot patterns and fixed chunks. Concrete lexical structures, frame-and-slot patterns, and fully abstract schemas are not mutually exclusive. All three types can coexist at different levels of representation in a hybrid exemplar-based categorization model. However, abstractness is always semantically restricted in the sense of usage-based approaches and can only be the result of a categorization process and not its beginning. Completely abstract linguistic schemata are thus always bound to more concrete ones or arise from them. This is linked to the idea that different levels of the degree of abstraction of linguistic representations are interconnected. Lexically concrete exemplars are in a vertical relationship with frame-and-slot patterns, whose abstraction basis is thus the concrete elements. However, the latter do not have to be lost after an abstraction process, but can remain mentally represented as well, depending on factors like their frequency of use. In this way, it is conceivable that the same linguistic pattern can be captured on different levels of a hybrid exemplar-based categorization model. For language acquisition, such a notion is highly relevant in that redundancies are allowed at the mental level. Lexically concrete constructions, as well as frame-and-slot patterns, do not have to be automatically lost as soon as an abstract pattern has been established. This is also relevant for the domain of code-mixing, where we are faced with the question of whether frequently recurring code-mixed patterns are just instantiations of more abstract constructions, or whether at least some of them also become entrenched as fixed chunks on repeated usage. The CBL can partly help answer this question—note, for example, that *noch mehr orange juice* is recognized as a chunk in



which both German and English material occurs. This could either be a coincidence that can be attributed to the distribution of orange juice in the data, or it could indicate that we are dealing with a pattern that has become entrenched, even though it has not occurred in the caregivers' data. This also points to a number of follow-up questions that could be addressed, e.g., to what extent a person's own language use contributes to entrenchment, rather than just the input that one receives.

**Author Contributions:** Conceptualization: N.K.; methodology, formal analysis, writing—original draft preparation, writing—review and editing, project administration: N.K., S.H. and A.E.Q.; Data curation: A.E.Q.; Software, visualization: S.H. All authors have read and agreed to the published version of the manuscript.

**Funding:** The open access publication of this article was funded by University and State Library (ULB) Düsseldorf.

**Institutional Review Board Statement:** Not applicable.

**Informed Consent Statement:** Informed consent was obtained from all participants involved in the study.

**Data Availability Statement:** The analysis script used for the present study can be found at <https://osf.io/zrkqm/>.

**Acknowledgments:** We would like to thank two anonymous reviewers for helpful comments and suggestions. Large parts of this paper were written during a research stay of the three authors at LMU Munich's Center for Advanced Studies (CAS), which was made possible by a "Researcher in Residence" fellowship awarded to the first author. We would like to thank the CAS for this opportunity and for providing the ideal environment for further developing our ideas in a highly stimulating and productive atmosphere.

**Conflicts of Interest:** The authors declare no conflict of interest.

## Notes

- <sup>1</sup> The age of the child is given in the format years;months or years;months.days. 3;8 thus means that the child is 3 years and 8 months old.
- <sup>2</sup> However, whether such a level actually exists mentally and, if so, is also used for the production and/or reception of language is currently being discussed intensively in usage-based approaches (see, e.g., [Abbot-Smith and Tomasello 2006](#); [Ambridge 2020a, 2020b](#)).
- <sup>3</sup> In TB studies, it is common to assume semantic restrictions for the slots in frame-and-slot patterns, which is why semantic categories are given here instead of, for example, part-of-speech or phrase categories.

## References

- Abbot-Smith, Kirsten, and Michael Tomasello. 2006. Exemplar-learning and schematization in a usage based account of syntactic acquisition. *Linguistic Review* 23: 275–90. [\[CrossRef\]](#)
- Adamou, Evangelia, and Yaron Matras, eds. 2020. *The Routledge Handbook of Language Contact*. London, New York: Routledge. [\[CrossRef\]](#)
- Ambridge, Ben. 2020a. Against stored abstractions: A radical exemplar model of language acquisition. *First Language* 40: 509–59. [\[CrossRef\]](#)
- Ambridge, Ben. 2020b. Abstractions made of exemplars or 'You're all right, and I've changed my mind': Response to commentators. *First Language* 40: 640–59. [\[CrossRef\]](#)
- Ambridge, Ben, and Elena Lieven. 2011. *Child Language Acquisition: Contrasting Theoretical Approaches*. Cambridge: Cambridge University Press. [\[CrossRef\]](#)
- Auer, Peter, and Carol Eastman. 2010. Code-switching. In *Society and Language Use*. Edited by Jürgen Jaspers, Jan-Ola Östman and Jef Verschueren. Amsterdam: John Benjamins, pp. 84–112. [\[CrossRef\]](#)
- Backus, Ad. 2014. Towards a Usage-Based account of language change: Implications of contact linguistics for linguistic theory. In *Questioning Language Contact: Limits of Contact, Contact at Its Limits (Brill Studies in Language Contact and Dynamics of Language 1)*. Edited by Robert Nicolai. Leiden and Boston: Brill, pp. 91–118. [\[CrossRef\]](#)
- Bannard, Colin, and Danielle Matthews. 2008. Stored Word Sequences in Language Learning. The Effect of Familiarity on Children's Repetition of Four-Word Combinations. *Psychological Science* 19: 241–48. [\[CrossRef\]](#) [\[PubMed\]](#)
- Bannard, Colin, Elena Lieven, and Michael Tomasello. 2009. Modeling children's early grammatical knowledge. *Proceedings of the National Academy of Sciences* 106: 17284–89. [\[CrossRef\]](#) [\[PubMed\]](#)
- Behrens, Heike. 2006. The input–output relationship in first language acquisition. *Language and Cognitive Processes* 21: 2–24. [\[CrossRef\]](#)

- Behrens, Heike. 2009. Usage-based and emergentist approaches to language acquisition. *Linguistics* 47: 383–411. [\[CrossRef\]](#)
- Belazi, Hedi M., Edward J. Rubin, and Almeida Jacqueline Toribio. 1994. Code Switching and X-Bar Theory: The Functional Head Constraint. *Linguistic Inquiry* 25: 221–37.
- Blom, Jan-Petter, and John Gumperz. 1972. Social meaning in linguistic structure: Code-switching in Norway. In *Directions in Sociolinguistics*. Edited by John Gumperz and Del Hymes. New York: Holt, Rinehart and Winston, pp. 407–34.
- Bybee, Joan L. 2010. *Language, Usage and Cognition*. Cambridge: Cambridge University Press. [\[CrossRef\]](#)
- Cameron-Faulkner, Thea, Elena Lieven, and Michael Tomasello. 2003. A construction based analysis of child directed speech. *Cognitive Science* 27: 843–73. [\[CrossRef\]](#)
- Cantone, Katja F. 2007. *Code-Switching in Bilingual Children*. Dordrecht: Springer.
- Chomsky, Noam. 1986. *Knowledge of Language*. Cambridge: MIT Press.
- Chomsky, Noam. 1987. Transformational grammar: Past, present, and future. In *Studies in English Language and Literature*. Kyoto: Kyoto University, pp. 33–80.
- Chomsky, Noam. 1995. *The Minimalist Program*. Cambridge: MIT Press.
- Clyne, Michael G. 1994. What can we learn from Sprachinseln? Some observations on ‘Australian German’. In *Sprachinselforschung. Eine Gedenkschrift für Hugo Jedig*. Edited by Nina Berend and Klaus Mattheier. Frankfurt: Lang, pp. 105–21.
- Clyne, Michael G. 2003. *Dynamics of Language Contact: English and Immigrant Languages*. Cambridge: Cambridge University Press.
- Crystal, David. 2003. *English as a Global Language*, 2nd ed. New York: Cambridge University Press. [\[CrossRef\]](#)
- Dąbrowska, Ewa. 2014. Recycling utterances: A speaker’s guide to sentence processing. *Cognitive Linguistics* 25: 615–53. [\[CrossRef\]](#)
- Dąbrowska, Ewa, and Elena Lieven. 2005. Towards a lexically specific grammar of children’s question constructions. *Cognitive Linguistics* 16: 437–74. [\[CrossRef\]](#)
- Di Sciullo, Anne-Marie, Pieter Muysken, and Rajendra Singh. 1986. Government and code-mixing. *Journal of Linguistics* 22: 1–24. [\[CrossRef\]](#)
- Goldberg, Adele E. 2006. *Constructions at Work*. Oxford: Oxford University Press. [\[CrossRef\]](#)
- Goldberg, Adele E. 2019. *Explain Me This. Creativity, Competition, and the Partial Productivity of Constructions*. Princeton: Princeton University Press. [\[CrossRef\]](#)
- Gumperz, John J. 1982. *Discourse Strategies. Studies in Interactional Sociolinguistics*. Cambridge: Cambridge University Press. [\[CrossRef\]](#)
- Gumperz, John J., and Dell H. Hymes. 1972. *Directions in Sociolinguistics: The Ethnography of Communication*. New York: Holt, Rinehart and Winston.
- Hartmann, Stefan, Nikolas Koch, and Antje Endesfelder Quick. 2021. The traceback method in child language acquisition research: Identifying patterns in early speech. *Language and Cognition* 13: 227–53. [\[CrossRef\]](#)
- Klann-Delius, Gisela. 2016. *Spracherwerb. Eine Einführung*, 3., aktual. u. erw. Aufl. Stuttgart: Metzler.
- Koch, Nikolas. 2019. *Schemata im Erstspracherwerb. Eine Traceback-Studie für das Deutsche*. Berlin: De Gruyter. [\[CrossRef\]](#)
- Koch, Nikolas, Stefan Hartmann, and Antje Endesfelder Quick. 2020. The traceback method and the early construction: Theoretical and methodological considerations. *Corpus Linguistics and Linguistic Theory*, 000010151520200045. [\[CrossRef\]](#)
- Kol, Sheli, Bracha Nir, and Shuly Wintner. 2014. Computational evaluation of the Traceback Method. *Journal of Child Language* 41: 176–99. [\[CrossRef\]](#) [\[PubMed\]](#)
- Lieven, Elena, Dorothé Salomo, and Michael Tomasello. 2009. Two-year-old children’s production of multiword utterances: A usage-based analysis. *Cognitive Linguistics* 20: 481–507. [\[CrossRef\]](#)
- Lieven, Elena, Heike Behrens, Jennifer Speares, and Michael Tomasello. 2003. Early syntactic creativity: A usage-based approach. *Journal of Child Language* 30: 333–70. [\[CrossRef\]](#) [\[PubMed\]](#)
- Lieven, Elena, Julian M. Pine, and Gillian Baldwin. 1997. Lexically-based learning and early grammatical development. *Journal of Child Language* 24: 187–219. [\[CrossRef\]](#)
- MacSwan, Jeff. 1999. *A Minimalist Approach to Intrasentential Code Switching*. New York: Routledge. [\[CrossRef\]](#)
- MacSwan, Jeff. 2000. The architecture of the bilingual language faculty: Evidence from intrasentential code switching. *Bilingualism: Language and Cognition* 3: 37–54. [\[CrossRef\]](#)
- MacSwan, Jeff. 2005. Codeswitching and generative grammar: A critique of the MLF model and some remarks on “modified minimalism”. *Bilingualism: Language and Cognition* 8: 1–22. [\[CrossRef\]](#)
- MacWhinney, Brian. 2000. *The CHILDES Project: Tools for Analyzing Talk*, 3rd ed. Hillsdale: Erlbaum.
- MacWhinney, Brian. 2004. A multiple process solution to the logical problem of language acquisition. *Journal of Child Language* 31: 883–914. [\[CrossRef\]](#) [\[PubMed\]](#)
- MacWhinney, Brian. 2014. Item-based patterns in early syntactic development. In *Constructions, collocations, patterns*. Edited by Thomas Herbst, Hans-Jörg Schmid and Susen Faulhaber. Berlin, Boston: De Gruyter, pp. 33–69. [\[CrossRef\]](#)
- McCauley, Stewart M., and Morten H. Christiansen. 2011. Learning simple statistics for language comprehension and production: The CAPPUCCINO model. *Proceedings of the Annual Meeting of the Cognitive Science Society* 33: 1069–7977.
- McCauley, Stewart M., and Morten H. Christiansen. 2014. Acquiring formulaic language: A computational model. *The Mental Lexicon* 9: 419–36. [\[CrossRef\]](#)
- McCauley, Stewart M., and Morten H. Christiansen. 2017. Computational Investigations of Multiword Chunks in Language Learning. *Topics in Cognitive Science* 9: 637–52. [\[CrossRef\]](#) [\[PubMed\]](#)
- McCauley, Stewart M., and Morten H. Christiansen. 2019. Language learning as language use: A cross-linguistic model of child language development. *Psychological Review* 126: 1–51. [\[CrossRef\]](#) [\[PubMed\]](#)

- McCauley, Stewart M., Padraic Monaghan, and Morten H. Christiansen. 2015. Language emergence in development. A computational perspective. In *Handbook of Language Emergence*. Edited by Brian MacWhinney and William O'Grady. Oxford: Blackwell, pp. 415–36. [\[CrossRef\]](#)
- Milroy, Lesley, and Pieter Muysken, eds. 1995. *One Speaker, Two Languages: Cross-Disciplinary Perspectives on Code-Switching*. Cambridge: Cambridge University Press. [\[CrossRef\]](#)
- Miorelli, Luca. 2017. The Development of Morpho-Syntactic Competence in Italian-Speaking Children: A Usage-Based Approach. Unpublished doctoral dissertation, Northumbria University, Newcastle upon Tyne, UK.
- Muysken, Peter. 2000. *Bilingual Speech: A Typology of Code-Mixing*. Cambridge: Cambridge University Press.
- Myers-Scotton, Carol. 1993. *Duelling Languages: Grammatical Structure in Codeswitching*. Oxford: Clarendon Press. [\[CrossRef\]](#)
- Myers-Scotton, Carol. 2006. *Multiple Voices: An Introduction to Bilingualism*. Malden: Blackwell Pub.
- Myers-Scotton, Carol, and Janice L. Jake. 2015. Cross-language asymmetries in code-switching patterns. Implications for bilingual language production. In *The Cambridge Handbook of Bilingual Processing*. Edited by John W. Schwieter. Cambridge: Cambridge University Press, pp. 416–58. [\[CrossRef\]](#)
- Paradis, Johanne, Elena Nicoladis, and Fred Genesee. 2000. Early emergence of structural constraints on code-mixing: Evidence from French+English bilingual children. *Bilingualism: Language and Cognition* 3: 245–61. [\[CrossRef\]](#)
- Pelucchi, Bruna, Jessica F. Hay, and Jenny R. Saffran. 2009. Learning in reverse: Eight-month-old infants track backward transitional probabilities. *Cognition* 113: 244–47. [\[CrossRef\]](#)
- Poplack, Shana. 1980. Sometimes I'll start a sentence in Spanish y termino en español. *Linguistics* 18: 581–618. [\[CrossRef\]](#)
- Quick, Antje Endesfelder, and Stefan Hartmann. 2021. The Building Blocks of Child Bilingual Code-Mixing: A Cross-Corpus Traceback Approach. *Frontiers in Psychology* 12: 682838. [\[CrossRef\]](#)
- Quick, Antje Endesfelder, Ad Backus, and Elena Lieven. 2018a. Partially schematic constructions as engines of development: Evidence from German-English bilingual acquisition. In *Cognitive Contact Linguistics*. Edited by Eline Zenner, Ad Backus and Esme Winter-Froemel. Berlin and Boston: De Gruyter, pp. 279–304. [\[CrossRef\]](#)
- Quick, Antje Endesfelder, Elena Lieven, Ad Backus, and Michael Tomasello. 2018b. Constructively combining languages: The use of code-mixing in German-English bilingual child language acquisition. *Linguistic Approaches to Bilingualism* 8: 393–409. [\[CrossRef\]](#)
- Quick, Antje Endesfelder, Elena Lieven, Malinda Carpenter, and Michael Tomasello. 2018c. Identifying partially schematic units in the code-mixing of an English and German speaking child. *Linguistic Approaches to Bilingualism* 8: 477–501. [\[CrossRef\]](#)
- Quick, Antje Endesfelder, Stefan Hartmann, Ad Backus, and Elena Lieven. 2021. Entrenchment and productivity: The role of input in the code-mixing of a German-English bilingual child. *Applied Linguistics Review* 12: 225–47. [\[CrossRef\]](#)
- Romaine, Suzanne. 1995. *Bilingualism*, 2nd ed. Oxford: Blackwell. [\[CrossRef\]](#)
- Romberg, Alexa R., and Jenny R. Saffran. 2010. Statistical learning and language acquisition. *WIREs Cognitive Science* 1: 906–14. [\[CrossRef\]](#) [\[PubMed\]](#)
- Rowland, Caroline F., Julian M. Pine, Elena V. M. Lieven, and Anna L. Theakston. 2003. Determinants of acquisition order in wh-questions: Re-evaluating the role of caregiver speech. *Journal of Child Language* 30: 609–35. [\[CrossRef\]](#) [\[PubMed\]](#)
- Sagae, Kenji, Brian MacWhinney, and Alon Lavie. 2004. Automatic parsing of parent-child interactions. *Behavior Research Methods, Instruments, and Computers* 36: 113–26. [\[CrossRef\]](#)
- Schmid, Hans-Jörg. 2017. A framework for understanding entrenchment and its psychological foundations. In *Entrenchment and the Psychology of Language Learning. How We Reorganize and Adapt Linguistic Knowledge*. Edited by Hans-Jörg Schmid. Berlin and Boston: De Gruyter, pp. 9–39. [\[CrossRef\]](#)
- Schmid, Hans-Jörg. 2020. *The Dynamics of the Linguistic System: Usage, Conventionalization, and Entrenchment*. Oxford: Oxford University Press. [\[CrossRef\]](#)
- Stoll, Sabine, Kirsten Abbot-Smith, and Elena Lieven. 2009. Lexically Restricted Utterances in Russian, German, and English Child-Directed Speech. *Cognitive Science* 33: 75–103. [\[CrossRef\]](#)
- Szagun, Gisela. 2019. *Sprachentwicklung Beim Kind*, 7th ed. Weinheim: Beltz.
- Tomasello, Michael. 2003. *Constructing a Language: A Usage-Based Theory of Language Acquisition*. Cambridge and London: Harvard University Press.
- Tomasello, Michael. 2015. The Usage-Based Theory of Language Acquisition. In *The Cambridge Handbook of Child Language*. Edited by Edith Laura Bavin and Letitia R. Naigles. Cambridge: Cambridge University Press, pp. 89–106. [\[CrossRef\]](#)
- Tomasello, Michael, and Patricia J. Brooks. 1999. Early syntactic development: A Construction Grammar approach. In *The Development of Language (Studies in Developmental Psychology)*. Edited by Martin Barrett. New York: Psychology Press, pp. 161–90.
- Vihman, Marilyn May. 1998. A Developmental Perspective on Codeswitching: Conversations Between a Pair of Bilingual Siblings. *International Journal of Bilingualism* 2: 45–84. [\[CrossRef\]](#)
- Vihman, Virve-Anneli. 2018. Language Interaction in Emergent Grammars: Morphology and Word Order in Bilingual Children's Code-Switching. *Languages* 3: 40. [\[CrossRef\]](#)
- Vogt, Paul, and Elena Lieven. 2010. Verifying theories of language acquisition using computer models of language evolution. *Adaptive Behavior* 18: 21–35. [\[CrossRef\]](#)
- Zabrodskaja, Anastassia. 2009. Evaluating the Matrix Language Frame model on the basis of a Russian—Estonian codeswitching corpus. *International Journal of Bilingualism* 13: 357–77. [\[CrossRef\]](#)